学校编号：10384　　分类号 _____密级 _____
学　　号：19120081152759　　UDC___

厦门大学

硕 士 学 位 论 文

系统进化网络的若干性质及度量

Phylogenetic networks and its metric

张 小 玲

指导教师姓名：钱 建 国 教授
专 业 名 称：应 用 数 学
论文提交日期：2011 年 5 月
论文答辩日期：2011 年 6 月
学位授予日期：2011 年 6 月

答辩委员会主席：_____
评　　阅　　人：_____

2011 年 5 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其它个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的权利和责任。

责任人（签名）：

年　月　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　　），在　　年解密后适用本授权书。

2、不保密（　　）。

（请在以上相应括号内打”✓”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

摘 要

系统网络作为系统发育树的推广，可以用来表示非类树的网状进化事件，如重组，杂交或后侧基因转移. 其基本思想是从现存的物种集所提供的信息构造出能够尽可能真实地反映物种进化过程的网络，在数学上可抽象为一个有根的有向无圈图. 一般而言，由于物种进化过程的多样性和复杂性，构造这样一个网络是极其复杂的. 例如，即使是构造一个具有最小杂交点数的系统网络也是一个 NP- 困难的问题 [3]. 因而，为了使问题更易于处理，研究者根据已知的进化规律对系统网络的构建提出了不同的限制和假设，并据此定义了各种系统网络，如：正则网络、正常网络、树孩子网络、galled-tree 、tree-sibling 网络及 tree-like 网络等. 特别地，根据任意两个物种的最近公共祖先总是存在的这一符合物种进化规律的假设， Willson [21] 定义了 mrca (most recent common ancestor)- 网络.

本文主要结果如下：

1. 研究了 mrca- 网络的性质并建立了与其他系统网络间的关系. 特别地，证明了 mrca- 网络严格地包含于正则网络，这样 mrca- 网络就具有了正则网络的性质及其上的度量.

2. 研究了系统网络的一些基本性质，给出了系统网络中的最大无关物种集的数目及现存物种集构成最大无关物种集的一个充分必要条件. 据此，证明了正常网络的叶子集构成了它的一个最大反链.

3. 作为 Robison-Foulds 度量的一个推广，证明了树孩子网络和半二叉 tree-sibling 时间连续性网络的 $\mu-$ 距离在满足杂交点没有立即的杂交孩子且不含 2 度树点的假设条件下也是正则网络上的一个度量.

关键词：系统网络，mrca- 网络，树孩子网络，正则网络，正常网络，最大反链，度量

Abstract

Phylogenetic networks are a generalization of phylogenetic trees that allow for the representation of non-treelike evolutionary events such as recombination, hybridization or lateral gene transfer. The basic idea is to construct a network from the information of extant taxa as a true reflection of species evolution. The network is typically modeled as an acyclic rooted directed graph. Generally speaking, due to the diversity and complexity of species evolution, constructing such a network is extremely complicated. Wang et al. [3] considered the problem of finding a perfect phylogenetic networks with the smallest recombination events and showed that the problem is NP-hard. Therefore, researchers proposed various restrictions on constructing a phylogenetic network in an attempt to make the problem more tractable, which yields a lot of phylogenetic networks, e.g., regular network, normal network, tree-child network, galled-tree, tree-sibling and tree-like network. Basing the reasonable assumption that the most recent common ancestor of any two species always exists, Willson [21] introduced the mrca-network.

The main results of this paper are as follows:

1. The properties of mrca-network are studied.
2. Some basic properties of phylogenetic networks related to the maximum antichain, the scale and the relationship among various networks are given.
3. As a generalization of Robison-Foulds' metric on tree-child networks and semibinary tree-sibling time consistent networks, the μ metric is shown to be valid for regular networks which satisfies some assumptions.

Keywords: Phylogenetic network, mrca-network, tree-child network, regular network, normal networks, maximum antichain, metric

目 录

Contents

记号

$d^+(u)$: u 的出度.

$d^-(u)$: u 的入度.

$p(u)$: u 的父亲集.

$child(u)$: u 的孩子集.

DAG: 无圈有向图.

$\mathrm{mrca}(W)$: W 的最近公共祖先.

$u \leq v$: 点 u 可通过一条有向路到达点 v.

$c(v) = \{x \in X | v \leq x\}$.

$C(N) = \{c(v) | v \in V\}$.

$V_L = \{l_1, \ldots, l_n\}$: 叶子集 (或现存物种集).

$C_L(u) = \{x | u \leq x, x \in V_L\}$.

$m_i(u)$: u 到叶子 l_i 的不同路的条数.

$\mu(u) = (m_1(u), \ldots, m_n(u))$.

$\mu(N) = \{\mu(u) | u \in V\}$.

第 1 章　引言

生物多样性的价值越来越被人们所认识与利用. 生物多样性的原因是生物进化的过程中, 物种和物种之间、物种和无机环境之间共同进化形成的结果. 对生物进行系统发育分析可以发现它们之间的亲缘关系及进化过程, 从而对其的利用更有针对性.

系统生物学是近年兴起的学科, 其创始人之一的美国科学家胡德 (Teroy Hood), 他说: “系统生物学将是二十一世纪医学和生物学的核心驱动力”. 近年来国内外很多大学和研究院纷纷成立系统生物学研究所或研究中心, 一些国际性的系统生物学研究会议也频繁召开. 那么什么是系统生物学呢? 据我国工程院院士杨胜利教授的定义, 系统生物学是“在细胞、组织、器官和生物体整体水平上研究结构和功能各异的生物分子及其相互作用, 并通过计算生物学来定量和预测生物功能、表型和行为”的这样一门学科. 系统生物学的理想就是要得到一个尽可能接近生物系统的理论模型; 建模过程贯穿于系统生物学研究的每个阶段. 离开了数学和计算机科学, 就不会有系统生物学. 系统生物学是适应于当前分子生物学的快速发展以及人类基因组计划等大科学工程而提出来的.

系统生物学不仅在医学和农学等方面具有重要的应用前景, 更重要的是它代表了生命科学理论的重大发展, 在生物进化的研究方向上就提出了许多新的发现. 如基因平行转移的发现、最小基因组与生命起源的关系、生物体可进化性概念的提出、无尺度网络生物模型的建立、蛋白质网络中与进化有关的各种 Hub 的发现等等. 这些新的概念和进化模型的提出, 都揭示了生物系统在分子相互作用下的演化规律, 并在不同程度上对其它层次上的演化具有非常重要的启发意义和应用价值.

自达尔文提出进化论以来, 人们普遍认为各种物种之间或多或少都存在某些亲缘关系. 科学家们根据进化论分析物种进化的规律并用这些规律进行

物种分类，种群及生物类群的演化研究. 然而，传统的进化分析只是从物种的形态，生活习性以及重要的指标进行分析，其间并不涉及分子水平下的进化研究. 随着现代生物学的发展，特别是基因测序以来，有了丰富的基因序列资源，科学家们更趋向于从序列上进行分子进化分析，这样的分析结果更能反映物种之间的亲缘关系. 尤其在揭示人类重大遗传疾病的分子基础、传染性疾病爆发与病原生物进化变异的关系，以及生物对环境变化的响应和适应机制方面显示出巨大的潜力，表现出显著的社会效应.

分子水平的进化研究致力于两大问题：重建物种间的进化关系以及了解进化过程的动力与机制. 重建物种间的进化关系属于系统学领域，传统上是用形态形状和化石来开展研究. 分子数据的广泛应用和简便易得已使其成为重建大部分物种类群系统发育关系中最常用的数据类型.

随着基因测序的发展，现在基因数据库里面的基因序列数以万计，其中最著名的基因数据库有 NCBI 、 EBI 、 DDBJ 数据库. 这说明基于序列分析的生物学时代已经到来. 生物信息学就是在如此庞大的基因序列数目下发展起来的. 生物信息学是把基因组 DNA 序列信息分析作为源头，然后根据序列信息应用数学与统计学方法计算出不同物种之间的同源序列差异，根据这些差异构建系统发育树.

生物信息学是系统发育研究中重要的分析手段，它通过 DNA 序列间的差异计算出核苷酸多态性，对于不同进化阶段物种的基因组结构和功能进行比较分析，可以追溯到一些基因的起源和进化过程，估算出生物之间的亲缘关系或遗传距离，并由此构建分子谱系树，推断群体的扩张模式、历史动态，推算群体起源、分歧的大致时间以及群体的进化速率、基因混合程度，甄别物种序列等，并可以给出统计学上的量化结果. 它可以从分子水平上探讨群体进化的规律，并可将这些规律以直观、形象的方式表现出来.

系统发育或系统发育树是物种间、基因间、群体间及至个体间谱系关系

的一种表现形式，是表达分类群之间系统发育关系的一种树状图. 其中，叶子代表今天的物种，内结点通常代表已灭绝祖先，树上的分枝（或边）代表发生基因突变. 它可以推测生物类群系统发育的分支样式，给出分支层次或拓扑图形，并能估算类群之间遗传关系的远近. 在生物进化研究中，通过构建系统发育树，可以推断个体之间以及群体间的亲缘关系，以及研究对象在系统树中所处的进化地位等.

但是，在某些植物和鱼类中，由于进化过程中，伴随发生杂交和水平基因转移等网状进化事件，致使许多物种的基因来自不同的祖先. 为了表示这种网状进化模型，人们提出了用系统网络 (有根的无圈有向图，简记为 DAG) 来表示物种进化关系. 其中，节点表示今天的物种和已灭绝的祖先，弧表示基因信息从一个物种转移到另一个物种. 至此，提出了许多有趣的数学和计算问题. 其中之一就是如何有效地表示和构造这些有向图. 由于在物种进化过程中，这种网状进化事件相对较少，因而对于这个问题，一个共同的方法是构造一个具有尽可能少网状事件的系统网络. 在文献 [18] 中， Wang 等考虑构造一类具有最少网状事件的完美系统网络，他们指出这个问题是一个 NP- 困难问题，一个具体的证明由 Bordewich 和 Semple [3] 给出.

因而，为了使问题更易于处理，研究者根据已知的进化规律对系统网络的构建提出了不同的限制和假设. 例如，文献 [10] 中 Wang 等考虑在系统网络中，重组事件相应于点不交的重组圈，同时给出构造这类网络的一个充分条件. 此后， Gusfield 等给出了构造这类网络的一个充分必要条件，并将这类网络称之为 galled-tree. 为了得到限制更弱的网络， galled-tree 的概念又进一步被推广到 level-k 网络 (每个双连通分支至少包含 k 个杂交点)[14]. 这样，根树是 level-0 网络而 galled-tree 是 level-1 网络. 在文献 [2, 4] 中，引进了正则网络 (即同构于它相应簇集的覆盖图) 和树孩子网络 (网络中的每个非叶子点均有一个正常孩子). 特别地，根据任意两个物种的最近公共祖先总

是存在的这一符合物种进化规律的假设， Willson[21] 定义了 mrca- 网络，这类网络的一个重要假设是不存在冗余弧，在不同的文献中，有些作者允许存在冗余弧，有些则不允许．例如，文献 [2] 中，由于正则网络同构于它簇集的覆盖图，为此，不允许有多余的弧．但在文献 [4] 中的树孩子网络允许存在冗余弧．

基于不同的系统网络，对任意两个不同物种的最近公共祖先存在的假设更符合物种进化规律．例如，假设某两个物种的最近公共祖先可追溯到一百万年前，我们就有理由推测这两个物种在一百万年前发生分歧．基于这个出发点，我们重新叙述了一类生物上合理的系统网络，称之为 mrca- 网络．本文，首先给出有关系统进化网络的一些基本概念．在第 3 和 4 节中，研究了 mrca- 网络的性质并建立了与其他系统网络间的关系．特别地，证明了 mrca- 网络严格地包含于正则网络，这样 mrca- 网络就具有了正则网络的性质及其上的度量．其次，研究了系统网络的一些基本性质，给出了系统网络中的最大无关物种集的数目及现存物种集构成最大无关物种集的一个充分必要条件．据此，证明了正常网络的叶子集构成了它的一个最大反链．

有时我们想度量两棵树之间有多少不同，如，我们可能对不同基因所估计的树之间的差异感兴趣，或者为了评价某种树重建方法而计算真实树和计算机模拟产生的估计树之间的差异．一种常见的两棵树之间的拓扑距离的测度是 Robinson-Foulds 定义的分划距离．同样的，对于相同的现存物种集 (叶子集)，基于不同系统网络的构造方法 [14, 16, 20-21]，可能得到不同的系统网络．因而，有必要找一个度量来比较这些系统网络．但迄今为止，对于一般系统网络，仍没有确切的度量，见 [5-6, 17]．另一方面，已知系统发育树的 Robinson-Foulds 度量可自然推广到正则网络．在文献 [4] 中，给出了树孩子网络的 $\mu-$ 度量．文章最后一部分证明了，虽然正则网络与树孩子网络不存在直接的关系，但树孩子网络上的 $\mu-$ 距离在满足杂交点没有立即的杂交孩

子且不含 2 度树点的假设条件下也唯一确定了正则网络. 由于 mrca- 网络严格包含于正则网络，因而，正则网络上的度量自然适用于 mrca- 网络.

总而言之，生物信息学就是一门预测性学科，根据已知的东西验证预测未知的东西. 对于系统发育树的展望，可以是根据已知序列的比对，找出各种物种之间特别是人类与其他物种之间的联系，以某种生物为研究对象来研究人类的各种生理生化的机理，让人们生活的更好. 例如在药物设计中，可以通过系统发育分析找出其他物种与人相近的同源序列，再以该物种为研究对象，研究药物时与靶点相作用的机理，从而避免了在人身上的直接实验，提高了安全性.

第 2 章　预备知识

一个有向图 $N=(V,E)$ 是由一个有限点集 V 和一个有限弧集 E 构成，其中 E 中每个元素是一个有序点对. 记 (u,v) 为 u 指向 v 的有向弧，并称 v 是 u 的孩子，u 是 v 的父亲. 若 u 和 v 有公共的父亲，则称 u (或 v) 是 v (或 u) 的兄弟姐妹. 用 $d^+(u)$ 与 $d^-(u)$ 分别表示 u 的出度与入度，出度为 0 的点称为 N 的叶子；出度大于 0 的点为内部点；入度为 0 的点为根点；入度小于等于 1 的点为树点 (或正常点); 入度大于 1 的点为杂交点. 点 u 的树孩子是指 u 的孩子且为树点.

一个有向图 N 的一条有向路是指一个点的序列 $v_1v_2\dots v_k$, 其中 $(v_i,v_{i+1})\in E, i=1,2,\dots,k-1$, 且序列中除 v_1 和 v_k 可能相同外，其它点均不相同. 定义有向路 $v_1v_2\dots v_k$ 的长为 $k-1$. 对于 $v\in V$, 以 v 为起点，叶子为终点的最长有向路的长称为点 v 的高度，并记为 $h(v)$. 特别地，当 $v_1=v_k$ 时，则称 $v_1v_2\dots v_k$ 是 N 的一个有向圈. 本文考虑的图均不含有向圈.

若点 u 可通过一条有向路到达 v, 则称 u 可达 v, 并记为 $u\le v$. 称一条弧 (u,v) 是冗余弧，如果存在不同于 u,v 的点 w 满足 $u\le w\le v$. V 的一个非空子集 W 的最近公共祖先 (most recent common ancestor) 是指满足下述条件的顶点 u (如果存在)，并记为 $\mathrm{mrca}(W)$:

(1) 对于任意的 $w\in W$, 均有 $u\le w$.

(2) 若 u' 也满足 (1), 则 $u'\le u$.

易看出，如果 $\mathrm{mrca}(W)$ 存在则它一定是唯一的 (若不然，设 u_1,u_2 为 W 的两个不同的最近公共祖先，则由 (2) 知，$u_1\le u_2$ 且 $u_2\le u_1$, 这与 N 不含有向圈矛盾).

一条有向路 $P=u_1u_2\dots u_k$ 称为基本路，如果 $d^+(u_i)=1, i=1,2,\dots,k-1$. 一条从 u 到 v 的有向路 $u=u_0,u_1,\dots,u_k=v$ 称为正常路，如果对于 $i=1,2,\dots,k$, u_i 均是正常点，这里 u 可能是正常点也可能是杂交点.

如果 u 到 X 中的某个点 x 有一条正常路，则称 u 到 X 有一条正常路．若 $u \in X$, 则 $u = u_0$ 是一条平凡的正常路．具有相同起点和终点的两条有向路构成一个圈 (看成无向圈), 称为重组圈．若一个重组圈与任意其他重组圈均无公共点，则称该重组圈是一个 gall.

一个系统网络 (如图 1) 定义为一个特殊的有叶子标号的有向图 $N = (V, E, r, X)$: V 和 E 分别是 N 的点集和弧集； r 是 N 的唯一的根； X 是 N 的基底集．没有杂交点的系统网络称为系统发育树．其中，基底集 X 表示基因信息可以通过直接测量而获得．如，叶子集代表现存的物种，故它们的 DNA(或基因) 信息可以直接检测得知．同时，根点通常被认为是一个遥远的祖先，在实践上可以用一个外类群代替．因而，也可以测得它的 DNA 信息．本文若无特殊声明我们均假设基底集 X 包含根，所有叶子及出度为 1 的点.

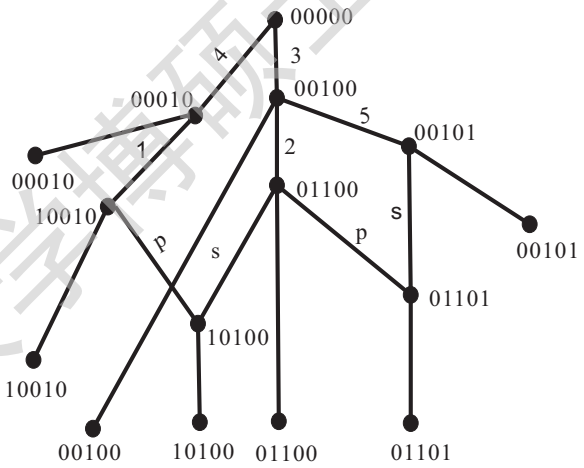

图 1: 从已知现存物种信息集 M={00010,10010,00100,10100,01100,01101,00101} 构造系统进化网络．其中，边上的数字标号 i 表示在 i 位置上发生基因突变．杂交点的基因信息由它的父亲集的基因信息的前缀和后缀重组所得，见 [10].

设 $P(X)$ 是 X 的某些子集构成的集族，文献 [2] 定义了一个簇映射 $c: V \to P(X)$, 其中 $c(v) = \{x \in X | v \leq x\}$ 称为 v 的簇集，记

$$C(N) = \{c(v) | v \in V\}.$$

设 $N = (V, E, r, X)$ 是一个系统网络，τ 是 V 到 N 的一个映射，满足：

(1) $\tau(r) = 0$.

(2) 若 v 是杂交点且 $(u, v) \in E$, 则 $\tau(u) = \tau(v)$.

(3) 若 v 是树点且 $(u, v) \in E$, 则 $\tau(u) < \tau(v)$.

则称 τ 是系统网络 N 上的一个时间分配. 若系统网络 N 存在时间分配，则称该网络 N 是时间连续的.

以下是本文涉及的一些特殊的系统网络 N.

正则网络 [2]: 系统网络 N 不含冗余弧且满足：

(1) 簇映射是一一映射. (2) $u \leq v$ 当且仅当 $c(v) \subseteq c(u)$.

正常网络 [19]: 系统网络 $N = (V, E, r, X)$ 的任意一个点 v 都有一条到 X 的正常路.

树孩子网络 [4]: 系统网络 N 的每个内部点都至少有一个树孩子.

galled-tree [10]: 系统网络 N 中的每个重组圈都是一个 gall, 见图 2.

mrca- 网络 [21]: 已知 $N = (V, E)$ 是一个系统网络，如果对于 V 的任意非空子集 U, $\mathrm{mrca}(U)$ 均存在.

tree-like 网络 [21]: 系统网络 $N = (V, E)$ 是一个 mrca- 网络且对于 V 中任意互异三点 x, y, z, 有 $\mathrm{mrca}(x, y, z) = \mathrm{mrca}(x, y)$ 或 $\mathrm{mrca}(x, y, z) = \mathrm{mrca}(x, z)$ 或 $\mathrm{mrca}(x, y, z) = \mathrm{mrca}(y, z)$.

tree-sibling 网络 [5]: 系统网络 N 中的每个杂交点都至少有一个兄弟姐妹是树点.

半二叉网络 [9]: 已知 N 是一个系统网络，如果对于 N 中的杂交点 u, 都有 $d^-(u) = 2$, 则称 N 是半二叉网络.